DOCUMENT RESUME

ED 385 576 TM 024 014

AUTHOR Pashley, Peter J.
TITLE Graphical IRT-Based DIF Analyses.
INSTITUTION Educational Testing Service, Princeton, N.J.
REPORT NO ETS-RR-92-66
PUB DATE Nov 92
NOTE 24p.
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Ability; Estimation (Mathematics); *Identification; *Item Bias; Item Response Theory; Psychometrics; *Statistical Analysis; *Test Items
IDENTIFIERS *Graphic Representation

ABSTRACT

The detection of differential item functioning (DIF) has become an important psychometric research topic in recent years. A number of item response theory (IRT) methods for solving this problem have been suggested. A common approach is to calculate some function of the area between item response curves estimated from the subpopulations of interest. While these methods relay overall item level DIF information, they do not indicate the location and magnitude of DIF along the ability continuum. In order to provide these important details, this paper presents a method for producing simultaneous confidence bands for the difference between item response curves. After these bands have been plotted, the size and regions of DIF are easily identified. Implementation considerations and illustrative examples are also given. One figure illustrates the discussion, and an appendix presents elements of information matrices associated with different parameters. (Contains 21 references.) (Author/SLD)

RR-92-66

RESEARCH REPORT

GRAPHICAL IRT-BASED DIF ANALYSES

Peter J. Pashley

Educational Testing Service
Princeton, New Jersey
November 1992

Graphical IRT-Based DIF Analyses

Peter J. Pashley

Educational Testing Service

Abstract

The detection of differential item functioning (DIF) has become an important psychometric research topic in recent years. A number of item response theory (IRT) methods for solving this problem have been suggested. A common approach is to calculate some function of the area between item response curves estimated from the subpopulations of interest. While these methods relay overall item level DIF information, they do not indicate the location and magnitude of DIF along the ability continuum. In order to provide these important details, this paper presents a method for producing simultaneous confidence bands for the difference between item response curves. After these bands have been plotted, the size and regions of DIF are easily identified. Implementation considerations and illustrative examples are also given.

Author Notes

Acknowledgements are gratefully extended to Charles Lewis and Rebecca Zwick for their comments and suggestions on the theoretical aspects of this paper; and to Howard Everson and Howard Wainer for reviewing the manuscript.

Graphical IRT-Based DIF Analyses

For many large and small standardized testing programs, checking for differential item functioning (DIF) has become a routine practice. This exercise exposes items that favor one subgroup over others due to characteristics that might be extraneous to the attributes being tested. With regard to this objective, impact studies (i.e., comparing average performance across groups) are insufficient unless average group performances are known to be equal, a priori. Preferred DIF methods partial out examinee abilities (or proficiencies) in some manner when comparing groups on a single item.

Two general approaches to this problem are usually taken. The first utilizes observed scores in DIF analyses. The most popular of this type is the Mantel-Haenszel (MH) procedure (Holland & Thayer, 1988), which is a chi-square type test. Within this procedure, so-called focal and reference groups are matched on an observed score, that may or may not include the item being investigated. Other observed score analyses that have been suggested include, the transformed item difficulty method (Angoff & Ford, 1973), the chi-square method (Camilli, 1979), and the standardization method (Dorans & Kulick, 1986).

The second general approach can be referred to as model-based analyses. The procedures that fall under this classification utilize examinee ability or true score estimates, and typically model the item attributes more extensively than those that use observed scores. Examples of these can be found in Cohen, Kim, and Subkoviak (1991); Lord (1977a, 1977b); and Thissen, Steinberg, and Wainer (1988). Comparisons of IRT with observed score methods can be found in Hambleton and Rogers (1989); Ironson and Subkoviak (1979);

Shepard, Camilli and Averill (1981); Shepard, Camilli and Williams (1985); and Subkoviak, Mack, Ironson and Craig (1984). One common approach is to calculate some function of the area between item characteristic curves (ICCs) estimated from the subpopulations of interest. [Formulas for computing the area between certain ICCs have been derived by Raju (1988)]. Unfortunately, this method yields only a single statistic and as such does not present an investigator with a clear picture of DIF over all parts of the ability range.

When entire ICCs have been employed (see, for example, Linn, Levine, Hastings and Wardrop, 1981; and Shepard, Camilli and Williams, 1984), only point-wise confidence bands have been used to account for sampling error, even though simultaneous confidence bands are usually more appropriate. In addition, the experiment-wise significance levels have not been properly controlled, in the past, to allow for sound statistical conclusions to be drawn. The purpose of this paper is to provide investigators with an IRT-based DIF analysis that utilizes simultaneous confidence bands for the difference between ICCs, and controls for the experiment-wise significance levels. After residual ICCs and their associated confidence bands are plotted, investigators are able to conveniently observe DIF regions and magnitudes.

The procedure presented here follows directly from the methodology for creating simultaneous confidence bands for single ICCs. With this in mind, an approach for calculating individual confidence bands will first be briefly discussed. Afterwards, a method for deriving confidence bands for the difference between two ICCs will be presented. Implementation considerations and examples are then discussed, followed by a summary.

Individual Simultaneous Confidence Bands

As mentioned in the introduction, the methodology for obtaining simultaneous confidence bands for the difference between two ICCs, follows directly from similar procedures for creating individual ICCs. Therefore, this basic methodology is presented here for completeness. Procedures for creating simultaneous confidence bands for logistic models whose logit is linear (i.e., the Rasch and two-parameter models) have been developed by Hauck (1983). This technique is very straightforward and will not be reproduced here. Note that the simultaneous confidence bands introduced by Hauck, and those presented here, assume that examinee proficiencies are known.

Individual Simultaneous Confidence Bands for the 3PL

The three-parameter logistic (3PL) does not possess a linear logit, and so Hauck's general approach can not be applied. The procedure presented here is similar to the Scheffé method for regression models [see, for example, Rao (1973), Sec. 4b.2]. The basic approach and final results are the same as those presented in Lord and Pashley (1988), although, some of the intermediate steps presented here are different. A more general discussion of this problem can be found in Thissen and Wainer (1990).

A common form of the 3PL, found in Lord (1980) and elsewhere, is given by

$$P(\theta) = c + \frac{1 - c}{1 + e^{-Da(\theta - b)}}, \qquad a > 0,\ 0 \leq c < 1,$$

$$= c + (1 - c)\Psi[Da(\theta - b)] ,$$

where $P(\theta)$ denotes the probability of correctly answering an item given an examinee ability

level θ; a, b, and c represent the discrimination, difficulty and lower asymptote item parameters, respectively; D is a constant, usually set to 1.7 or 1.702; and $\Psi(\cdot)$ is the logistic function.

The form of the 3PL used in this paper is given by

$$P(\theta) = c + (1 - c)\Psi(A\theta + B) ,$$

where $A\theta + B$ is simply a reformulation of $Da(\theta - b)$. Maximum likelihood estimates (MLEs) of A and B can be obtained from the MLEs of a and b, due to their invariance property, as follows:

$$\hat{A} = D\hat{a}, \text{ and } \hat{B} = -D\hat{a}\hat{b}.$$

To simplify notation, let the vectors $\beta = (A, B, c)'$ and $\hat{\beta} = (\hat{A}, \hat{B}, \hat{c})'$ contain the unknown item parameters and their corresponding MLEs, respectively, associated with a single item. We assume that β was estimated from a sample of known θ_i's ($i = 1, 2, \ldots, N$) whose properties do not preclude the usual asymptotic normality assumptions associated with the distribution of $\hat{\beta}$. In particular, we assume that with sufficiently large N, the quadratic form

$$(\hat{\beta} - \beta)'\Sigma^{-1}(\hat{\beta} - \beta) \doteq \chi^2_{(3)} , \qquad (1)$$

where Σ is the asymptotic covariance matrix associated with $\hat{\beta}$; $\doteq$ reads "is approximately

distributed as"; and $\chi^2_{(3)}$ denotes the central chi-square distribution with three degrees of freedom.

As is commonly done, the expected sample information matrix, denoted **I**, will be used as an estimate of Σ^{-1}. One could also employ the observed sample information matrix, though this alternative may be less well-behaved under certain circumstances. After substituting **I** into (1), we can define a $1 - \alpha$ confidence ellipsoid for β by

$$\text{Prob}[(\beta - \hat{\beta})' \mathbf{I} (\beta - \hat{\beta}) \leq \chi^2_{(3,\alpha)}] = 1 - \alpha , \qquad (2)$$

where $\chi^2_{(3,\alpha)}$ denotes the upper α percentage point of the $\chi^2_{(3)}$ distribution. Note that since the quadratic form is only "approximately distributed" as chi-square, the inequality in (2) and all related inequalities and equalities that follow, actually only denote asymptotic approximations.

Now let the three-dimensional constraint space for β, defined by (2), be denoted by S. Then for a fixed θ, we can map S into a one-dimensional constraint space for $P(\theta)$, which can be represented by the interval

$$\{\min[P(\theta) | \beta \in S],\ \max[P(\theta) | \beta \in S]\} .$$

Then using a repeated sampling argument, these intervals will define a $1 - \alpha$ simultaneous confidence band for $P(\theta)$, over all θ. What remains to be done is to provide a procedure for obtaining, for any fixed θ, the endpoints of these intervals.

The task of finding the maxima and minima of $P(\theta)$, for a fixed ϑ and given a constraint space S, can be formulated in non-linear programming terms. The general problem can be stated as

$$\begin{aligned} &\text{optimize} && P(A,\ B,\ c;\ \theta) \\ &\text{subject to} && (\hat{\beta} - \beta)' \mathbf{I} (\hat{\beta} - \beta) \leq \chi^2_{(3,\alpha)} , \end{aligned}$$

where the notation $P(A,\ B,\ c;\ \theta)$ is used to emphasize which parameters will be varied in order to obtain an optimal (i.e., maximum or minimum) value of $P(\theta)$.

In order to simplify this problem, a further reparameterization is useful. In particular, we let

$$L = A\theta + B .$$

Then

$$P(A,\ B,\ c;\ \theta) = P(L,\ c;\ \theta) = c + (1 - c)\Psi(L) .$$

The constraint space can be similarly reparameterized, and at the same time reduced from an ellipsoid to an ellipse on the $L \times c$ plane, by using the linear transformation/reduction matrix

$$\mathbf{M} = \begin{bmatrix} \theta & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}$$

as follows:

$$(\hat{\beta} - \beta)' \mathbf{I} (\hat{\beta} - \beta) \geq (\hat{\beta} - \beta)' \mathbf{M} (\mathbf{M}' \mathbf{I}^{-1} \mathbf{M})^{-1} \mathbf{M}' (\hat{\beta} - \beta) = (\hat{\delta} - \delta)' \mathbf{J} (\hat{\delta} - \delta) ,$$

where

$$\delta = \mathbf{M}'\beta = (L, c)' ,$$
$$\hat{\delta} = \mathbf{M}'\hat{\beta} = (\hat{L}, \hat{c})' , \text{ and}$$
$$\mathbf{J} = (\mathbf{M}' \mathbf{I}^{-1} \mathbf{M})^{-1} .$$

In addition, note that the first partial derivatives of the objective function,

$$\frac{\partial P(L, c; \theta)}{\partial L} = \frac{1 - c}{(1 + e^{L})(1 + e^{-L})} , \text{ and}$$

$$\frac{\partial P(L, c; \theta)}{\partial c} = \frac{1}{1 + e^{L}}$$

are always positive. Hence, the optimal values of $P(L, c; \theta)$, for a fixed θ, will be found on the boundary of the constraint ellipse. The reparameterized and simplified optimization problem is then

$$\text{optimize} \quad P(L, c; \theta)$$
$$\text{subject to} \quad (\hat{\delta} - \delta)' \mathbf{J} (\hat{\delta} - \delta) = \chi^2_{(3,\alpha)} .$$

After solving the quadratic constraint equation for L and evaluating appropriate ranges for c, the problem of finding a lower bound can be stated as

$$\text{minimize} \quad c + (1 - c)\Psi(L)$$

$$\text{subject to} \quad L = \hat{L} + \frac{(\hat{c} - c)J_{Lc} - \sqrt{J_{LL}\chi^2_{(3,\alpha)} - (\hat{c} - c)^2(J_{LL}J_{cc} - J^2_{Lc})}}{J_{LL}}$$

$$\hat{c} - \sqrt{\frac{J_{LL}\chi^2_{(3,\alpha)}}{J_{LL}J_{cc} - J^2_{Lc}}} \leq c \leq \hat{c} + \frac{J_{Lc}}{J_{cc}}\sqrt{\frac{J_{cc}\chi^2_{(3,\alpha)}}{J_{LL}J_{cc} - J^2_{Lc}}} .$$

Similarly, the problem of determining the upper bound can be formulated as

$$\text{maximize} \quad c + (1 - c)\Psi(L)$$

$$\text{subject to} \quad L = \hat{L} + \frac{(\hat{c} - c)J_{Lc} + \sqrt{J_{LL}\chi^2_{(3,\alpha)} - (\hat{c} - c)^2(J_{LL}J_{cc} - J^2_{Lc})}}{J_{LL}}$$

$$\hat{c} - \frac{J_{Lc}}{J_{cc}}\sqrt{\frac{J_{cc}\chi^2_{(3,\alpha)}}{J_{LL}J_{cc} - J^2_{Lc}}} \leq c \leq \hat{c} + \sqrt{\frac{J_{LL}\chi^2_{(3,\alpha)}}{J_{LL}J_{cc} - J^2_{Lc}}} .$$

Unfortunately, as $P(L, c; \theta)$ does not constitute a convex set, multiple local maxima and minima are possible. However, two line searches for the maxima and minima can easily be conducted by varying c between the extreme values indicated. Formulae for calculating the J_{ij}'s are given in the Appendix.

Residual Simultaneous Confidence Bands

We will proceed in a fashion similar to the one taken in the previous section. First let the vector $\lambda = (A_F, B_F, c_F, A_R, B_R, c_R)'$ contain the item parameters corresponding to the focal (F) and reference (R) groups (i.e., the two groups of interest). Then after making the usual normality assumptions, the residual optimization problem can be expressed as:

$$\text{optimize} \quad P(A_F, B_F, c_F; \theta) - P(A_R, B_R, c_F; \theta)$$

$$\text{subject to} \quad \lambda' \Sigma^{-1}_{(\lambda)} \lambda \leq \chi^2_{(6,\alpha)} .$$

We now evoke the assumption of local item independence. This implies that the between-group item parameter covariances are all equal to zero. The above constraints can then be written as

$$\beta_F' \Sigma^{-1}_{(\beta_F)} \beta_F + \beta_R' \Sigma^{-1}_{(\beta_R)} \beta_R \leq \chi^2_{(6,\alpha)} ,$$

where $\beta_F = (A_F, B_F, c_F)'$ and $\beta_R = (A_R, B_R, c_R)'$.

With this insight, the original residual optimization problem can be rewritten as

$$\text{optimize} \quad P(A_F, B_F, c_F; \theta) - P(A_R, B_R, c_F; \theta)$$

$$\text{subject to} \quad \beta_F' \Sigma^{-1}_{(\beta_F)} \beta_F \leq \gamma \chi^2_{(6,\alpha)}$$
$$\beta_R' \Sigma^{-1}_{(\beta_F)} \beta_R \leq (1 - \gamma) \chi^2_{(6,\alpha)}$$
$$0 \leq \gamma \leq 1 .$$

Written in this fashion, the residual optimization problem can now be undertaken as a series of individual simultaneous confidence band problems. This is achieved in practice by increasing γ incrementally, for a fixed θ, and recording the maximum and minimum differences.

Implementation Considerations and Examples

Various sampling and calibration procedures have been suggested for IRT-based DIF analyses in the past. The methodology presented in the previous section should conform to most of these procedures. Only one method, however, was used to produce the examples that follow. This approach is comprised of the following four basic steps:

1) Select representative samples from focal and reference groups.
2) Calibrate all examinees together on a set of non-DIF items.
3) Calibrate items of interest, separately by group, using ability estimates obtained in Step 2.
4) Calculate and plot residual ICCs and corresponding simultaneous confidence bands.

Note that while matched samples are not required, obtaining representative samples from the subgroups of interest should, in most cases, be very important.

Examples

Items from a large-scale Educational Testing Service administered examination were investigated. The focal and reference groups of interest were female and male examinees, respectively. The samples were comprised of approximately 1,250 females and 1,500 males. All examinees were first calibrated together on 98 operational (assumed to be non-DIF) items to obtain proficiencies on a common scale. Holding these values fixed, experimental items were then calibrated, separately, for the two groups. The results for three experimental items are presented here. These three items had previously been classified, using different samples, as "A", "B", and "C" items, based on a MH analysis. These three classifications,

"A", "B" and "C", refer to low, medium and high levels of DIF, respectively.

Residual ICCs and associated simultaneous 95% confidence bands, for each of the three items, are given in Figure 1, alongside corresponding individual ICCs with labeled ranges containing evidence of DIF . The results are clearly in agreement with the previous MH analysis. The simultaneous confidence band for the first ("A") item completely encompasses the zero-residual reference line, indicating no evidence of DIF along the entire range of proficiencies. The band for the second ("B") item dips slightly below the reference line, within the .2 to 1.4 range of abilities. The third ("C") item's band drops significantly below the reference line, between the ability values -1.2 and 1.8.

Discussion and Conclusions

Since some subgroups are not always well represented across the entire ability scale, matching samples, as required by most observed score analyses, can be a problem. As evident in the previous section, the calibration samples used in the graphical IRT-based approach need not be matched. In addition, as seen in the Figure 1, the location and magnitude of the DIF along the ability scale is easily appreciated with the proposed approach.

The three example items given confirmed the results of a previous MH analysis. In general, this is exactly what one would hope to find. However, as with certain other IRT approaches, the method presented here allows for the possibility of uncovering bi-directional DIF, where the corresponding ICCs cross each other. In these cases, the focal group

performs better than the reference group across some parts of the ability continuum, but worse across other segments. These "plus" and "minus" DIF regions can cancel each other out within a MH analysis.

The method presented in this paper utilizes simultaneous confidence bands for the difference between ICCs. One may ask whether there are occasions where point-wise bands would be more appropriate. In cases where only a few points along the ability scale are of interest, calculating simultaneous bands could be excessive. For example, certification tests may possess one or two cut-scores of special interest. In these cases point-wise confidence intervals may actually be preferable. Otherwise, if a large segment of proficiency scale is of interest, calculating simultaneous confidence bands would usually be more appropriate.

While this proposed procedure maintains the trappings of a significance test, it should be regarded as an exploratory data analysis technique, as all sources of error are not accounted for. In particular, the sampling errors related to the estimation of examinee abilities are not included in the calculation of the simultaneous confidence bands. In any case, the DIF effect size should be viewed as the most important aspect to consider. The proposed method simply tempers enthusiasm for seemingly significant effect sizes by clearly illustrating associated sampling variation.

References

Angoff, W. H., & Ford, A. F. (1973). Item-race interaction on a test of scholastic aptitude. Journal of Educational Measurement, 10, 95-105.

Camilli, G. (1979). A critique of the chi-square method for assessing item bias. Unpublished manuscript, University of Colorado-Boulder, Laboratory of Educational Research.

Cohen, A. S., Kim, S-H, & Subkoviak, M. J. (1991). Influence of prior distributions on detection of DIF. Journal of Educational Measurement, 28, 49-59.

Dorans, N. J., & Kulick, E. (1983). Assessing unexpected differential item performance of female candidates on SAT and TSWE forms administered in December 1977: An application of the standardization approach (RR-83-9). Princeton, NJ: Educational Testing Service.

Hambleton, R. K., & Rogers, H. J. (1989). Detecting potentially biased test items: Comparison of IRT area and Mantel-Haenszel methods. Applied Measurement in Education, 2 313-334.

Hauck, W. W. (1983). A note on confidence bands for the logistic response curve. The American Statistician, 37, 158-160.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. Braun (Eds.), Test validity (Pps. 129-145). Hillsdale, NJ: Lawrence Erlbaum Associates.

Ironson, G. H., & Subkoviak, M. J. (1979). A comparison of several methods of assessing

item bias. Journal of Educational Measurement, 16, 209-225.

Linn, R. L., Levine, M. V., Hastings, C. N., & Wardrop, J. L. (1981). Item bias in a test of reading comprehension. Applied Psychological Measurement, 5, 159-173.

Lord, F. M. (1977a). Practical Applications of item characteristic curve theory. Journal of Educational Measurement, 14, 117-138.

Lord, F. M. (1977b). A study of item bias using item characteristic curve theory. In Y. H. Poortingal (Ed.), Basic problems in cross-cultural psychology. Amsterdam: Swets & Zeitlinger.

Lord, F. M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Erlbaum.

Lord, F. M., & Pashley, P. J. (1988). Confidence bands for the three-parameter logistic item response curve (Research Report RR-88-67). Princeton, NJ: Educational Testing Service.

Raju, N. S. (1988). The area between two item characteristic curves. Psychometrika, 53, 495-502.

Rao, C. R. (1973). Linear statistical inference and its applications (2nd ed.), New York: Wiley.

Shepard, L. A., Camilli, G., & Averill, M. (1981). Comparison of procedures for detecting test-item bias with both internal and external ability criteria. Journal of Educational Statistics, 4, 317-375.

Shepard, L. A., Camilli, G., & Williams, D. M. (1984). Accounting for statistical artifacts

in item bias research. Journal of Educational Statistics, 9, 93-128.

Shepard, L. A., Camilli, G., & Williams, D. M. (1985). Validity of approximation techniques for detecting item bias. Journal of Educational Measurement, 22, 77-105.

Subkoviak, M. J., Mack, J. S., Ironson, G. H., & Craig, R. D. (1984). Empirical comparison of selected item bias detection procedures with bias manipulation. Journal of Educational Measurement, 21, 49-58.

Thissen, D., Steinberg, L., & Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. In H. wainer & H. Braun (Eds.), Test validity, (pp. 147-169). Hillsdale, NJ: Lawrence Erlbaum Associates.

Thissen, D., & Wainer, H. (1990). Confidence envelopes for item response theory. Journal of Educational Statistics, 15, 113-128.

Appendix

The elements (J_{ij}) of the information matrix **J** associated with the parameters L and c, can be expressed in terms of the elements (I_{ij}) from the information matrix associated with the parameters A, B and c, as follows:

$$J_{LL} = I_{BB} - \frac{(I_{AB} - \theta I_{BB})^2}{I_{AA} - 2\theta I_{AB} + \theta^2 I_{BB}}$$

$$J_{Lc} = I_{Bc} - \frac{(I_{AB} - \theta I_{BB})(I_{Ac} - \theta I_{Bc})}{I_{AA} - 2\theta I_{AB} + \theta^2 I_{BB}}$$

$$J_{cc} = I_{cc} - \frac{(I_{Ac} - \theta I_{Bc})^2}{I_{AA} - 2\theta I_{AB} + \theta^2 I_{BB}}$$

Lord and Pashley (1988) reported the elements of the information matrix associated with the parameters A, B, and c, in terms of the elements of the usual information matrix for a, b, and c (Lord, 1980, p. 191), as follows:

$$I_{AA} = \frac{I_{aa} + \frac{b}{a}\left[\frac{b}{a}I_{bb} - 2I_{ab}\right]}{D^2}$$

$$I_{AB} = -\frac{I_{ab} - \frac{b}{a}I_{bb}}{D^2 a}$$

$$I_{BB} = \frac{I_{bb}}{D^2 a^2}$$

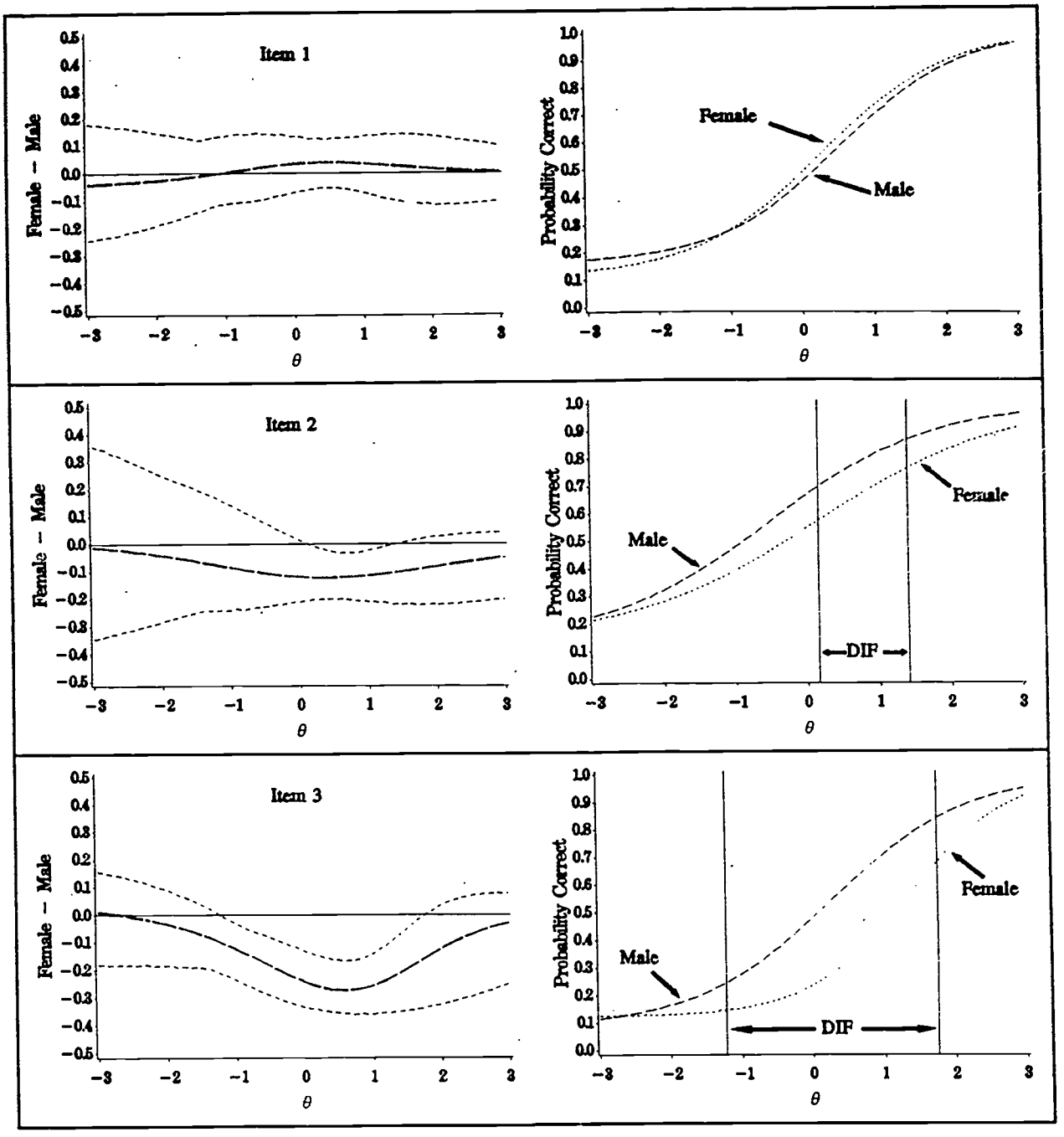
$$I_{Ac} = \frac{I_{ac} - \frac{b}{a}I_{bc}}{D}$$

$$I_{Bc} = -\frac{I_{bc}}{Da}$$

The element I_{cc} remains the same under these two parameter definitions.

Figure Caption

Figure 1: Graphical IRT-based DIF results for three items. Each panel contains (1) a residual plot with an associated 95% simultaneous confidence band, and (2) a graph illustrating the ICCs pertaining to the focal and reference groups. The regions labeled "DIF" refer to the ranges of θ for which there is evidence of DIF.


Item 1
Female − Male
Probability Correct
Female
Male
θ
Item 2
Female − Male
Probability Correct
Male
Female
DIF
θ
Item 3
Female − Male
Probability Correct
Male
Female
DIF
θ